arXiv:2210.02925v1 [cs.FL] 6 Oct 2022

Yet another proof of Parikh's Theorem

Manfred Kufleitner

University of Stuttgart, FMI, Germany
kufleitner@fmi.uni-stuttgart.de

Abstract. Parikh's Theorem says that the Parikh image of a context-free language is semilinear. We give a short proof of Parikh's Theorem using the formulation of Verma, Seidl, and Schwentick in terms of Presburger arithmetic. The proof relies on an Eulerian property of derivation trees of context-free languages and was inspired by Hierholzer's algorithm; it does not use the Chomsky normal form.

These notes are based on a presentation given in Bordeaux on June 16th, 2015 at a workshop in honor of Volker Diekert's 60th birthday.

A *context-free grammar* $G = (N, \Sigma, P, S)$ consist of a finite set of *non-terminals* N, a finite alphabet Σ, a finite set of *productions* $P \subseteq N \times (N \cup \Sigma)^*$, and a *start symbol* S. The elements of Σ are also called *terminals* and the productions are also known as *rules*. The rules are usually written as $X \to u$ for $X \in N$ and $u \in (N \cup \Sigma)^*$; they induce a string rewrite system on $(N \cup \Sigma)^*$: For $v, w \in (N \cup \Sigma)^*$ we write $v \Rightarrow w$ if $v = pXq$ and $w = puq$ for $p, q \in (N \cup \Sigma)^*$ and a rule $X \to u$ in P. We consider the reflexive and transitive closure $\overset{*}{\Rightarrow}$ of the relation $\Rightarrow$. If $v \overset{*}{\Rightarrow} w$, then w is a *derivation* of v. The language $L(G)$ generated by G consists of all derivations of the start symbol S which do not contain any non-terminal:

$$L(G) = \left\{ w \in \Sigma^* \;\middle|\; S \overset{*}{\Rightarrow} w \right\}$$

A language is *context-free* if it is generated by a context-free grammar.

A *derivation tree* is an ordered tree t with labeled nodes. The label of the root is the start-symbol S, the labels for the inner nodes come from the non-terminals N, all leafs have a label in $\Sigma \cup \{\varepsilon\}$ where ε denotes the empty word, and if an inner node has label X, then the sequence of labels of its children forms a word $u \in (N \cup \Sigma)^*$ such that $X \to u$ is in P. If the sequence of labels of the leafs forms a word w, then we say that t is a derivation tree for w. We have $S \overset{*}{\Rightarrow} w$ if and only if there exists a derivation tree for w. There can be more than one derivation tree for w.

Let $|w|_a$ denote the number of occurrences of the letter a in the word $w \in \Sigma^*$. The *Parikh image* of w is

$$\Psi(w) = (|w|_a)_{a \in \Sigma}$$

As usual, for $L \subseteq \Sigma^*$ we set $\Psi(L) = \{ \Psi(w) \mid w \in L \}$. If $\Sigma = \{a_1, \ldots, a_k\}$, then the Parikh mapping $\Psi : \Sigma^* \to \mathbb{N}^k$ is the homomorphism defined by $\Psi(a_i) = (0, \ldots, 0, 1, 0, \ldots, 0)$ with a one at the i-th component; here, the operation on $\mathbb{N}^k$ is component-wise addition.

A subset of the form $\{ u + x_1 v_1 + \cdots + x_n v_n \mid x_i \in \mathbb{N} \} \subseteq \mathbb{N}^k$ for $u, v_i \in \mathbb{N}^k$ is called *linear*. A *semilinear* set is a finite union of linear sets. Parikh's Theorem says that Parikh images

of context-free languages are semilinear. We equivalently state Parikh's Theorem in terms of Presburger arithmetic, the first-order logic of $(\mathbb{N}, +)$. A formula in Presburger arithmetic is *existential* if there is no universal quantifier and no negation. We use the proof of Verma, Seidl, and Schwentick [5] to construct a Presburger formula for a given context-free grammer G. While [5] relies on a result of Esparza [1] for showing that every model of the formula has a corresponding word in $L(G)$, we give a new self-contained proof for this part.

Theorem 1 (Verma, Seidl, Schwentick [5]) *Given a context-free grammar G, one can compute an existential Presburger formula for the Parikh image of $L(G)$ in linear time.*

Proof: Let $G = (N, \Sigma, P, S)$ with $|\Sigma| = k$. For a rule $p = (Y \to u)$ and $X \in N \cup \Sigma$ we define

$$X(p) = \begin{cases} 1 & \text{if } X = Y \\ 0 & \text{otherwise} \end{cases}$$

and we set $|p|_X = |u|_X$. We define a Presburger formula φ for $\Psi\big(L(G)\big)$. For every rule $p \in P$ we introduce a variable x_p which represents the number of applications of p. For $X \in N \setminus \{S\}$ we define

$$\alpha_X := \left(\sum_{p \in P} X(p) \cdot x_p = \sum_{p \in P} |p|_X \cdot x_p \right)$$

and

$$\alpha_S := \left(\sum_{p \in P} S(p) \cdot x_p = 1 + \sum_{p \in P} |p|_S \cdot x_p \right)$$

The formula α_X says that the number of occurrences of X on the left side of a rule application matches the number of X's on the right side of a rule application. Since S is the start symbol, there is one more occurrence on the left than on the right. For every non-terminal $X \in N$ we introduce an index y_X for numbering consecutively the first occurrences of the non-terminals in the derivation tree. We ensure that $y_X = 0$ if and only if X does not occur. Let $\beta_S := (y_S = 1)$ and for $X \neq S$ we define

$$\beta_X := \left(y_X = 0 \ \wedge \sum_{p \in P} |p|_X \cdot x_p = 0 \right) \vee \bigvee_{\substack{(Y \to u) \in P \\ |u|_X > 0}} \big(x_{Y \to u} > 0 \ \wedge \ y_Y > 0 \ \wedge \ y_X = y_Y + 1 \big)$$

The formula φ has a free variable z_a for the number of occurrences of the letter $a \in \Sigma$. We define

$$\gamma_a := \left(z_a = \sum_{p \in P} |p|_a \cdot x_p \right)$$

The formula φ is the conjunction of all these formulas, with existentially quantified variables x_p and y_X, i.e., for $\overline{x} = (x_p)_{p \in P}$ and $\overline{y} = (y_X)_{X \in N}$ we set

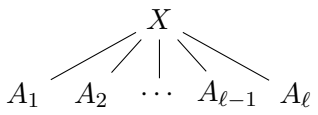
$$\varphi \ := \ \exists \overline{x} \, \exists \overline{y} \colon \bigwedge_{X \in N} (\alpha_X \ \wedge \ \beta_X) \ \wedge \bigwedge_{a \in \Sigma} \gamma_a$$

If $w \in L(G)$, then there exists a derivation tree for w. Let x_p be the number of occurrences of the rule $p \in P$ in this tree. Every inner node corresponds to both an occurrence on the left side of a rule and an occurrence on the right side of a rule, the sole exception is the root S. Thus $\overline{x} \models \bigwedge_{X \in N} \alpha_X$. We define y_X inductively. Let $y_S = 1$ and if y_X is not yet defined but for some parent Y of X the value y_Y is already defined, then we set $y_X = y_Y + 1$. Note that all non-terminals X in the derivation tree satisfy $y_X > 0$ since there cannot be a topmost undefined non-terminal X with a previously defined parent Y. All ultimately undefined variables y_X are set to zero. Whenever we define $y_X > 0$ for $X \neq S$, then this is due to some rule $Y \to u$ with $|u|_X > 0$ for some previously defined variable y_Y; in particular $x_{Y \to u} > 0$. Therefore $\overline{x}, \overline{y} \models \bigwedge_{X \in N} \beta_X$. Let $z_a = |w|_a$ and $\overline{z} = (z_a)_{a \in \Sigma}$. Every letter a of w is a leaf of the derivation tree, and every leaf in the derivation tree originates from some rule. This yields $\overline{x}, \overline{z} \models \bigwedge_{a \in \Sigma} \gamma_a$. We conclude $\overline{z} \models \varphi$ and $\Psi\big(L(G)\big) \subseteq L(\varphi)$.

For the remaining inclusion $L(\varphi) \subseteq \Psi\big(L(G)\big)$ we show that if $(z_a)_{a \in \Sigma} \models \varphi$, then there exists a derivation tree for a word w with $|w|_a = z_a$ for all $a \in \Sigma$. Let $\overline{z} = (z_a)_{a \in \Sigma} \in \mathbb{N}^k$ with $\overline{z} \models \varphi$. Then there exist numbers $\overline{x} = (x_p)_{p \in P}$ and $\overline{y} = (y_X)_{X \in N}$ with

$$\overline{x}, \overline{y}, \overline{z} \models \bigwedge_{X \in N} (\alpha_X \wedge \beta_X) \wedge \bigwedge_{a \in \Sigma} \gamma_a$$

We take x_p copies of the rule p. If p is of the form $X \to A_1 \cdots A_\ell$ with $A_i \in N \cup \Sigma$, then we represent p as a tree:



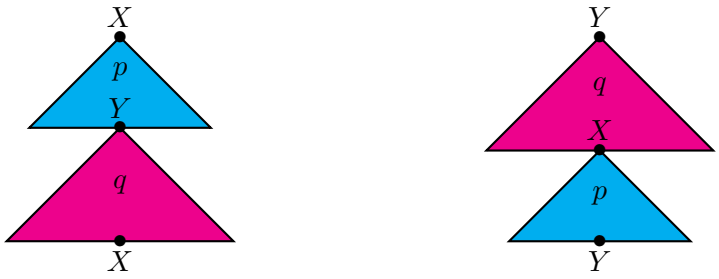
tree for $\ell > 0$

tree for $\ell = 0$

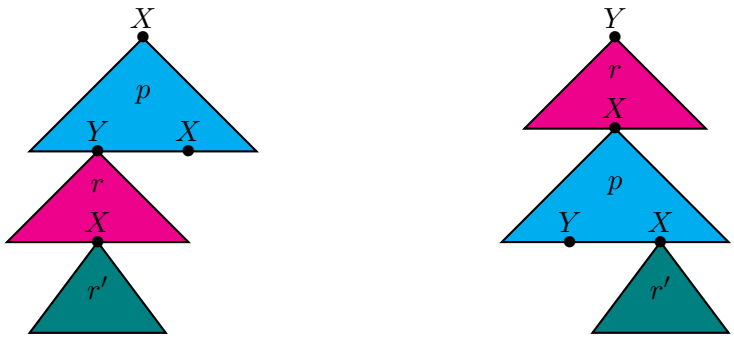
Let r_X be the number of occurrences of X as a root in all these trees, and let s_X be the number of leafs with label X. The formula α_X ensures $r_X = s_X$ for $X \neq S$ and $r_S = s_S + 1$. The formula γ_a says that there are z_a leafs with label a. We iterate the following procedure as long as possible: If there is a tree t_1 with leaf X and another tree t_2 with root X, the trees t_1 and t_2 are replaced with a new tree $t_1 \cdot t_2$ by identifying an X-leaf of t_1 with the root of t_2. This reduces r_X by 1 and s_X by 1. Thus throughout we have $r_X = s_X$ for $X \neq S$ and $r_S = s_S + 1$. The number of leafs with label $a \in \Sigma$ remains unchanged. Eventually, there cannot be two trees with root X since then there has to be some tree with leaf X, which in turn could then be combined with one of the trees with root X into a new tree. Thus we end up with a tree with root S in which all leafs are terminals and with a set of trees of the following form: The root is $X \in N \setminus \{S\}$ and all leafs are terminals except for one leaf which is X.

Suppose that we have a tree t with root X and leaf X. If there exists another tree s (possibly the tree with root S) with some node labeled by X, then we can factorize $s = p \cdot q$ at the non-terminal X. We replace the two trees s and t by the tree $p \cdot t \cdot q$. This reduces the number of trees by 1.

Otherwise, all occurrences of X are inside t. By definition of β_X, this also includes a subtree for a rule $Y \to u$ with $|u|_X > 0$ and $y_X = y_Y + 1$. If the non-terminal Y of this subtree is on the path from the root X to the leaf X, then we can factorize $t = p \cdot q$ at Y and replace t by $q \cdot p$.



Otherwise, we can factorize $t = p \cdot q$ at this particular occurrence of Y. Since u defines the child nodes of the root Y in q and since $|u|_X > 0$, we can factorize $q = r \cdot r'$ at X. Now, we can replace t by $r \cdot p \cdot r'$.



In any case, this reduces the index of this root by 1. By repeating the previous steps as long as possible, eventually there is only one tree left. This is a derivation tree for a word w with $|w|_a = z_a$ for all $a \in \Sigma$.

The formula φ can be computed in linear time since for instance in $\sum_{p \in P} X(p) \cdot x_p$ and $\sum_{p \in P} |p|_X \cdot x_p$ it suffices to consider those rules p with $X(p) \neq 0$ and $|p|_X \neq 0$, respectively. □

Corollary 2 (Parikh [3, 4]) *If $L \subseteq \Sigma^*$ is context-free, then $\Psi(L) \subseteq \mathbb{N}^{|\Sigma|}$ is semilinear.*

Proof: If L is context-free, then $\Psi(L)$ is definable in Presburger arithmetic by Theorem 1. Ginsburg and Spanier [2] showed that Presburger definable subsets of $\mathbb{N}^k$ are semilinear. □

References

[1] J. Esparza. Petri nets, commutative context-free grammars, and basic parallel processes. *Fundam. Informaticae*, 31(1):13–25, 1997.

[2] S. Ginsburg and E. H. Spanier. Semigroups, Presburger formulas, and languages. *Pacific Journal of Mathematics*, 16(2):285–296, 1966.

[3] R. J. Parikh. Language generating devices. *Quarterly Progress Report*, 60:199–212, 1961.

[4] R. J. Parikh. On context-free languages. *Journal of the ACM*, 13:570–581, 1966.

[5] K. N. Verma, H. Seidl, and T. Schwentick. On the complexity of equational horn clauses. In *CADE-20, Proceedings*, volume 3632 of *LNCS*, pages 337–352. Springer, 2005.